IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Volume 13, PP 43-45

A Study of Translator Camera Using Optical Character Recognition & Natural Language Processing.

Aditya Prabhu, Mihir Pitroda, Susmit Wadikar, Prof. Suvarna Pansambal
Department of Computer Engineering Atharva College of Engineering, Malad (W) Mumbai, India.

Abstract : *Systems today are constantly evolving to achieve accurate & efficient results day by day. The sole purpose is to help human in their day to day queries. Today Artificial Intelligence (AI) is present in a variety of fields ranging from trivial guidance mechanisms, Virtual Assistants, Navigation Guides, to large scale industrial co-ordination systems. There are numerous AI assistants that help people solve their problems. This paper is a summarization of the overall research conducted to achieve the principled goal of digital image data translation, for the targeted content of Language. The paper summarizes a brief overview on the existing systems and algorithms that may be efficient in the process of Language Translation. As countries that have a huge cultural diversity, tend to have multiple set of languages that are used locally or nation-wide, this creates a communication gap due to existence of a language barrier. Therefore, it is a matter of priority to enable the systems to develop or adapt to techniques that ease this barrier.*

Keywords: *OCR- Optical Character Recognition; NLP- Natural Language Processing, Image Processing, Artificial Intelligence, TDM-Transform Domain Maps, SVM-Support Vector Machine, NLC- Natural Language Computing, UNL-Universal Networking Language.*

I. Introduction

Language Translation is basically analysis of a source language as an input that is processed and converted to an almost literal sentence formation in the desired target output, with the primary goal to preserve the exact meaning. The analysis conducted in this field covers the various Algorithmic discoveries, existing systems, and other implemented ideologies.

In recent years, recognition and translation of text written on display boards has become hot topic of research in computer vision. Automatic text recognition and translation is useful in various applications such as: tour guide system, robotic servers in hotels, blind assistants and many more.

II. Need

India has the world's second highest number of languages. This creates a communication gap for people visiting the regions with different languages. This can be significantly observed where the signboards and notices are written in unfamiliar languages which are unable to be interpreted for non-locals.

According to Census of India of 2001, India has 122 major languages and 1599 other languages. Hence there is a major communication gap between the localities of different regions especially if the languages are in written form, where even interpreting it using pronunciation is not an option. This scenario leads to even more problems for tourists who originate from foreign countries. All of this leads to them being scammed and misguided by some localities which effectively creates a bad reputation of country and negatively impacts the tourism industry.

This all can be avoided if there is an autonomous system which can help the user to translate the written material from local language to language of their understanding. Effectively eliminating the involvement of some middle-man and a degree of dependency.

III. Literature Survey

Language translation mechanism that we have taken under study includes the basic principles of OCR & NLP of the domains of Artificial Intelligence & Image processing.

R. R. Karnik [2] identified the Devnagri characters using OCR. As most of the languages in India are partial or completely based on devanagri script. The problem of identifying a devanagri script is that, most of the words are a conglomeration of multiple syllables and most of the script is cursive in nature, therefore, script identification in OCR becomes tedious. Prominent script practices like Hindi, Marathi, Sanskrit and Konkani visually have a similar appearance with minor differences for identification. Therefore the OCR extraction done on such inputs is a prima facie task of impossibilities. Because of the simplistic reason that in the devanagri script if characters from the word were separable there would be more than thousand patterns making it very hard to identify the ideal solution of the correct meaning. Another striking feature of the devanagri script is that not only the characters are written under a line to form a word there are also characters like the kana or matra

that are written above, below or inline. There are certain vowels or vowel consonants that are written below the words.

Even though the process of OCR has been discussed there exists the issue of the output accuracy and efficiency of the algorithm. Certain factors of the input image turn out to be a hindrance in the effective and in time generation of results. For e.g. - For a text input, missing strokes of the character or misprinted character or smudges on the surface which are reasons for incorrect interpretations by the OCR engine. Therefore Naveen Sankaran [8] proposed Bidirectional Long Short Term Memory (BLSTM) approach for character recognition. This approach reduces the word error rate by 20%. and 9% in character error rate.

There exists a different approach to get a highly accurate and robust OCR system will also work as Handwritten Optical Character Recognition (HOCR) allowing handwriting analysis as a part of the system. As the accuracy of OCR depends on three primary levels: line, character segmentation, word. Therefore the use of support vector machines and Transform Domain Maps improve the overall accuracy by the implementation of close point symmetry. Even though the implementation process of SVM and TDM is able to generate multi class classification, it still compromises certain extent of speed for gaining accuracy [4].

Xin Jia [9] used another approach of the OCR engine by assigning AI agent trained in Deep Learning algorithms. Deep learning algorithm strategy allows the agent to gather more abstract information making the generation of the output easier to understand. It also improves the generalization ability of the engine making a significant increase in output generation.

The extracted data of the OCR engine cannot be directly translated. Therefore a local repository with source and destination language data should be made available. An effective approach for such a repository is to use multiple variants of the same word structure, thus increasing the hit ratio of the OCR query. As a result, drastically improves the speed and accuracy of the system [1].

An alternative way for a multi directional translation base repository is the lexicon approach to be used. This format structure uses lexical reference system constructed to store the variant sentences of the language. For every information of any word, category, punctuation, adjective, noun, etc. the storage in repository would be in the form of a lexical code making it possible for an effective two way translation process[7].

The translation process used is a fully automated hybrid based machine translation. This translation method uses a large amount of parallel corpora enabling the system to use dictionary sentences, bilingual dictionary and phonetic dictionary based approaches to produce the translation along with a set of rules descriptively forming a sentential identification structure making translation of literal accurate and with the objective of keeping the original meaning intact [3].

Along with the above permutation of the text identification, extraction and translation factors we can also implement another methodology to obtain similar results. This system can produce the proper translations by the use of proper techniques such as corpus management, multilingual lexical database and Natural Language Computing (NLC) [6].

To ensure and improve the efficiency of the desired system a practical test approach where samples of input across multiple individuals should be gathered and fed to the system and conducting analysis and development of statistics on key factors like intelligence, accuracy and speed. Depending on the classification of the test cases that are fed to the system [5].

IV. Available Mechanisms

Most widely used translation systems take typed text as input; hence one has to be able to write the words of the language, which is a major problem for someone who is not familiar at all.

1.1. Anusaarkaa

An Indian machine translation program from Telugu, Kannada, Bengali, Punjabi & Marathi as a source language to Hindi as a destination language or vice versa was developed in 1995 by the name of Anusaarkaa. It works on the principles of Paninian Grammar which is primarily used to generalize the constituents and replace them from extracted form of the raw example [6].

1.2. Anubharti- II

This technology is a system which implements a combination of example based, corpus based and some elementary grammatical analysis approaches. The traditional approach of translation in Anubharti (predecessor) has been modified to replace the large scale example based requirement by the Anubharti-II in 2004 [6] [3].

1.3. UNL based English-Hindi machine translation

A machine translation system that was successfully able to translate from English to Hindi using UNL as Interlingua was developed at IIT Bombay in 2003. UNL was an international project of the United Nations University with the goal to create an Interlingua for all major languages [6].

1.4. Sampark

A machine translation mechanism for Indian languages was proposed by a consortium of multiple Indian Institutes (IIT Hyderabad, University of Hyderabad, C-DAC Noida, Anna University Bangalore, IIT Ahmedabad, Tamil Nadu University, and Jadhavpur University). A set of languages including Punjabi, Hindi, Telgu, Tamil was proposed in 2005.

1.5. The Mantra (Machine Translation Tool)

A Machine Translator , introduced in 1999, which gives precise English to Hindi and vice-versa translation, on the basis of administrative directives for Office Orders in precise accuracy.

In the current generation there is not an app or program which can do the task of extracting the text from an image and then converting it to preferred language by the user simultaneously. There exist different algorithms to implement each task individually. This is our effort to combine these two tasks together and make an efficient system which can be beneficial for a lot of different activities. There are different algorithms which perform best individually but when combined together to do both the task simultaneously then the complexity is increased or time is increased or accuracy is reduced or the database to recognize the text and translate is too much. Hence, In order to provide an efficient app that does both the task efficiently and gives the best results, these two methods and combination of algorithms fitted best when combined together.

V. Conclusion

The language translation system will help overcome the simplest yet the major difficulty faced by many people who visit or reside in an area of which they know very little or nothing about. Even if there is a communication gap between people then this help may come in handy to bridge that gap by using this powerful tool. As this is system evolves it may become very easily accessible as most of the people that possess a cellular device with a camera functionality. This is the system will help in easing the language barrier in communication.

References

[1] Archana Singh, Raj Shree, "Recognition of Natural Language Processing to Manage Digital Electronic Applications", Volume 8, No. 5, May-June 2017 International Journal of Advanced Research in Computer Science.
[2] R. R. Karnik, "Identifying Devnagri Characters", Volume 6, No. 12, May-June 2015,
[3] Sindhu.D.V , Sagar.B.M, "Study on Machine translation approaches for Indian languages and their challenges",2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)
[4] Saurabh Farkya, Govinda Surampudi, and Ashwin Kothari, "Hindi Speech Synthesis by concatenation of recognized Hand written Devnagri Script using Support Vector Machines Classifier", IEEE ICCSP 2015 conference.
[5] Vishal Goyal and Gurpreet Singh Lehal, "Evaluation of Hindi to Punjabi Machine Translation System", IJCSI International Journal of Computer Science Issues, Vol. 4, No. 1, 2009.
[6] Mrs. Shachi Mall, Dr. Umesh Chandra Jaiswal, "Developing a System for Machine Translation from Hindi language to English language", 2013 4th International Conference on Computer and Communication Technology (ICCCT).
[7] Mallikarjun M. Kodabagi, S. A. Angadi, "A Methodology for Machine Translation of Simple Sentences from Kannada to English Language"
[8] Naveen Sankaran, C.V Jawahar, "Recognition of Printed Devanagari Text Using BLSTM Neural Network", 21st International Conference on Pattern Recognition (ICPR 2012)
[9] Xin Jia "Image Recognition Method Based on Deep Learning"